On scoring rules and frequency predictive measures

Ksenia N. Kyzyurova
Sheffield, UK

January 29, 2019

Abstract

Scoring rules of statistical models are getting popular among statisticians and practitioners for model comparison and model selection. We provide pedagogical examples with beautiful geometric interpretation and illustrations which demonstrate that scoring rules have very limited ability in doing so and are capable of misguiding on predictive model assessment and comparison. Predictive model comparison is demonstrated to be subjective. We recommend to, instead, employ three independent measures of model predictive performance: (1) empirical frequency coverage, (2) an estimate of predictive bias and (3) an estimate of uncertainty (variability) in predictions.

Keywords: Model assessment, model comparison

1 Introduction

Scoring rules are mathematical procedures which compare how close a predictive distribution, given by a particular model, is to a true value observed in nature or experiment. A rule assigns a value of the score to a model based on this comparison. See Gneiting & Raftery (2007). This paper aims at alarming statisticians and practitioners who utilize scoring rules for model assessment and model comparison that the ability of scores in doing so is very limited.

Indeed, problematic case studies have been reported before, and date back to at least 1970's. See, for example, Murphy & Winkler (1970) who raise concerns on the applicability of scoring rules in practice. Another interesting and more recent example is given in Smith et al. (2015) who provides a clear illustration of problems with model assessment with continuous ranked probability score (CRPS).

In particular this work grew out from research on emulation of computer models with multivariate output considered in dissertation of Kyzyurova (2017). Analysis involved model comparison with multiple scoring rules.

For brevity of exposition we define a score as a mathematical procedure which assigns a numerical value to a predictive distribution p with respect to the observed value x. Mathematical details and a survey of scores may be found in Gneiting & Raftery (2007), Dawid et al. (2015). In the rest of the paper we use logarithmic score because of its popularity. The score has a correspondence to likelihood-ratio test and Bayes factor, thus allowing for better interpretability of this score Kass & Raftery (1995) as opposed to many other scores.

Despite of these interpretations, the log-score (the same way as other scores) suffers from the problem that evaluated scores do not form a totally ordered set, and thus comparison

of competing models becomes a challenging problem. In its turn implementation of the scores in practical applications leads to poorly interpretable results, which may potentially be detrimental to the consequences of such model assessment.

2 Problem formulation and illustration

Let p be a predictive density of a forecast given by some model for a random variable in a one-dimensional space R_1. A positively oriented log-score which evaluates this distribution by comparison to the true value x is as follows

$$\text{logS}(p, x) = \log p(x), \tag{1}$$

Consider a class of normal predictive distributions. Then for $p = \mathcal{N}(\mu, \sigma)$, normal distribution with mean μ and standard deviation σ, the score with respect to an observed value x is

$$\text{logS}(p, x) = -\frac{\log(2\pi\sigma^2)}{2} - \frac{(x-\mu)^2}{2\sigma^2}\,. \tag{2}$$

Assuming that the true value x equals to zero, a plot with contourlines of scores which in this case depend simply on the mean and standard deviation of a distribution p is shown on Figure 1. The contourlines show that (1) since there are infinitely many points along a particular contourline, infinitely many distributions exist which have exactly the same value of a log-score.

The same value of a score make these distributions to be considered equally good predictive distributions, but (2) these distributions may be very different from each other. As example, consider two points on the plot: the green point corresponds to a predictive distribution $\mathcal{N}(\mu = -1.7, \sigma = 1.038363)$ and orange point corresponds to a distribution

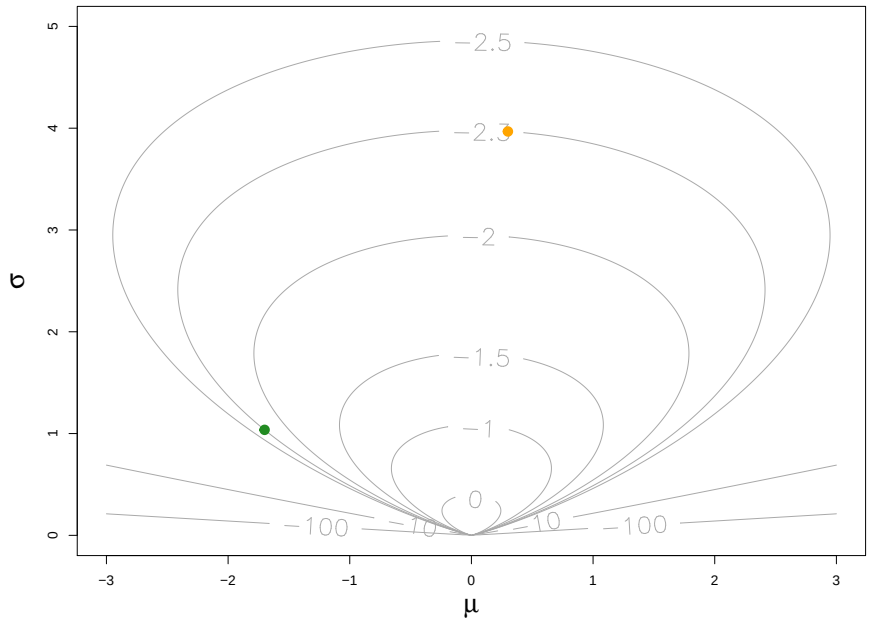

Figure 1: Contourlines of the logarithmic score given for a class of normal distributions characterized by two parameters: mean μ and standard deviation σ. Two coloured points correspond to two distributions.

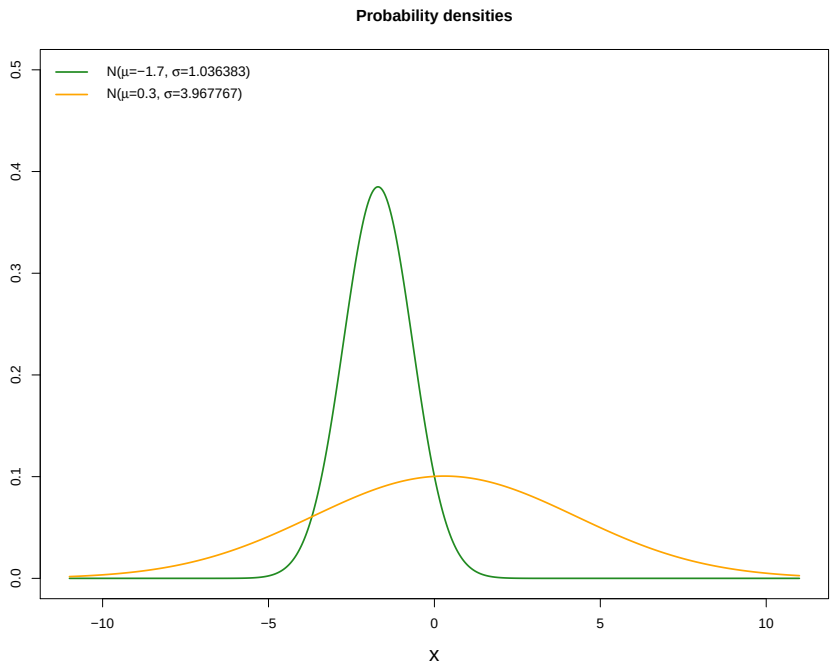

Figure 2: Two distributions corresponding to points of the same colour in Figure 1. These two distributions have exactly the same value of a logarithmic score.

$\mathcal{N}(\mu = 0.3, \sigma = 3.967767)$. Probability densities of these distributions are shown in Figure 2. Indeed, the distributions are very different: the green distribution has a large bias — the mean of the distribution is far from the true value of zero. The orange distribution has bias more than 5 times smaller of that of the green one. However, the orange distribution has standard deviation more than three times greater than that of the green one, which translates to variance being more than 14 times greater. Yet, two distributions are assigned exactly the same value of a log-score.

Consider another example of two similar distributions which could have resulted from

two other models, namely $\mathcal{N}(\mu = 0.9, \sigma = 1)$ and $\mathcal{N}(\mu = 1, \sigma = 0.9)$. The plot of densities of corresponding distributions is given in the supplementary materials. Providing virtually the same information on the predictive distribution in practice, the scores, as compared to the true value of zero, are -1.323939 and -1.430862 respectively, thus differentiating between the two distributions much more than one would expect.

These examples demonstrate that logarithmic score can not distinguish between very different distributions, assigning the same scores to considerably different shapes; and at the same time assigning much more different value to very similar distributions. It may be demonstrated that other scores would provide similar but different contourplots for the class of normal distributions, behaving in analogous way as have been observed with the logarithmic scoring rule. Additional example with illustrations of such behaviour using seven scoring rules is given in the supplementary materials. That example demonstrates that considered scoring rules result in different preferences of the models and their predictive distributions.

3 Predictive measures

Yet, the comparison of a models' performance given by different predictive distributions is desired. There may be two possibilities for a solution to this problem: (1) subjective judgement and (2) attempt to employ a more-or-less objective judgement relying on frequency performance of a predictive distribution.

Subjective choice does not require much discussion. For instance, one person may choose a model which gives a predictive distribution with smaller bias. Another one may choose a model which gives a distribution with smaller variance. In this sense any scoring rule can be viewed as someone's subjective evaluation criteria with respect to predictive

distributions.

Objective choice has been proved to be useful, for instance, for assessment of statistical models' quality in literature on Uncertainty Quantification. See examples in, e.g. Gu & Berger (2016), Kyzyurova et al. (2018). Namely, three predictive measures are employed. First measure is empirical coverage, which states if the true point belongs to a central 95% area of the predictive distribution or not.[1][2] Second measure is bias, the discrepancy between, e.g. the mean of the distribution and the true value $|\mu - x|$. Third measure is the length of a 95% credible interval given by a distribution. These measures are intuitively preferred over scores because they provide clear interpretation of what they mean in the evaluation of a predictive performance of a model.

Since we assess a distribution using its probabilistic credible area, frequency evaluation of this distribution needs to be employed. Say, m observations from experimental data are available $x_1, \ldots, x_m$, then three frequency performance measures are formulated as follows: the proportion of true values captured by a predictive distribution within its central 95%-credible interval.

$$\text{EFC} = \left(\sum_{i=1}^{m} I_{x_i \in \text{CI}_\text{i}} / m \right) 100\% \in (0, 100)\,, \tag{3}$$

root-mean-square predictive error

$$\text{RMSPE} = \sqrt{\sum_{i=1}^{m} (x_i - \mu_i)^2 / m} \in (0, \infty)\,, \tag{4}$$

and the average length of the 95% credible intervals over m points. For normal predictive

[1] In our experience 95% credible area has served as a useful nominal value. However, we acknowledge that this choice is subjective. Other nominal values may be employed.

[2] We choose to work with *central* 95% credible area because of its invariance under transformations as opposed to highest posterior density (HPD) areas.

distribution this is defined as

$$\overline{\mathrm{L_{CI}}} = \sum_{i=1}^{m} 2 \cdot 1.96\sigma_i/m \in (0, \infty)\,. \tag{5}$$

We recommend to consider these three measures independently, without collapsing them into one measure. The latter attempt has been shown to lead to an *improper* score. See Gneiting & Raftery (2007) discussion on predictive model choice criterion (PMCC) and the meaning of (im)propriety of a score.

An ideal case of when predictive distribution coincides or almost coincides with the true distribution, and 95% of experimental observations should fall within 95% central credible area of the distribution. In practice exact numbers of 95% are rarely achievable. One needs to subjectively assess if, say 80% empirical coverage versus corresponding 95% area is good enough. For other two predictive measures, It is desirable that RMSPE and $\overline{\mathrm{L_{CI}}}$ are small. If several models are considered, then ratios of the corresponding measures are useful for assessment of how much one model is better than another in each of these measures.

Interestingly, an attempt to perform frequency assessment with the scores results in their impropriety, allowing to generate additional criticism for this choice.

4 Conclusion

In this work we have shown that while very modest model comparison via scoring rules is possible, the characterization of the differences among competing models is typically not possible. Once the values of the scores of two models are sufficiently similar, comparison of these models using *any* score becomes a virtually impossible task.

When one model is substantially better than another in all three predictive measures, empirical frequency verification is enough. However, one needs to be prepared to a situation

when the choice between two distributions needs to rely on a subjective preference. For example, if number of evaluation points m is low for frequency assessment or if predictive distributions are close to each other so that it is hard to judge which one is better.

In the present work we have only considered a class of normal distributions. We could have chosen any other class of distributions. Arguably, normal distribution appears more often than other distributions in both, pedagogical examples and applications. We believe this choice assisted with construction of simple examples which provide clear illustrations to the main message of this manuscript.

SUPPLEMENTARY MATERIAL

Title: Pdf-file contains a complementary figure for the example in the main part of the manuscript and additional example with illustrations and numerical summaries with respect to six more scoring rules as well as logarithmic score.

References

Dawid, A. P., Musio, M. et al. (2015), 'Bayesian model selection based on proper scoring rules', *Bayesian analysis* **10**(2), 479–499.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Gu, M. & Berger, J. O. (2016), 'Parallel partial gaussian process emulation for computer models with massive output', *The Annals of Applied Statistics* **10**(3), 1317–1347.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Kyzyurova, K. N. (2017), On Uncertainty Quantification for Systems of Computer Models, PhD thesis, Duke University.

Kyzyurova, K. N., Berger, J. O. & Wolpert, R. L. (2018), 'Coupling computer models through linking their statistical emulators', *SIAM/ASA Journal on Uncertainty Quantification* **6**(3), 1151–1171.

Murphy, A. H. & Winkler, R. L. (1970), 'Scoring rules in probability assessment and evaluation', *Acta psychologica* **34**, 273–286.

Smith, L. A., Suckling, E. B., Thompson, E. L., Maynard, T. & Du, H. (2015), 'Towards improving the framework for probabilistic forecast evaluation', *Climatic Change* **132**(1), 31–45.